

GOexplore & GOcompare: Shiny apps to visualize Gene Ontology Annotations

Gaurav Kandoi

May 4, 2017

Abstract

A fundamental aim of genetics is to understand the functions of proteins which is critical for understanding the molecular basis of diseases and other phenotypic traits. The Gene Ontology (GO) Consortium provides controlled vocabularies of defined terms representing gene product (protein) properties. These GO terms assigned to genes as annotations describe three aspects of the gene: 1) Cellular Component; 2) Molecular Function and; 3) Biological Process. The gene annotations assign GO terms to the genes along with an evidence code that depict the strength of the annotation. Some evidence codes are more reliable than others, for instance, annotations derived from experimental evidence codes are more valid than those derived from automatic analyses. A typical gene annotation data provides association details for thousands of genes and GO terms with the aspect, evidence codes, literature citation, date and assigned by among others. The evidence code, aspect, gene and GO terms are multi-factor variables making them a perfect choice for exploring the quality of the data using their combinations. Because there can be multiple evidences supporting the relation between a gene and a GO term, the annotations are redundant. I'll write functions to remove the redundancy and other inherent biases in the annotations data. The aim of this project is to develop a shiny app to provide an intuitive access to the gene annotation data by means of providing descriptive visualization. Since the GO annotations are regularly updated, the app will always fetch the latest annotations when invoked to provide analysis using the current annotations. The platform will also provide an easy to use interface for comparing the quality of annotations across species.

Gene Ontology

The GO project is an extensive resource for functional genomics. It provides evidence-supported annotations for gene products using ontologies to describe the biological roles of these gene products. GO is structured as a Directed Acyclic Graph (DAG), where every node represents a GO term which have defined relationships with one or more terms. GO describes three aspects of the gene products:

- Cellular Component- cellular locations where these occur
- Molecular Function- molecular level elemental activities
- Biological Process- series of one or more molecular events